DOCUMENT RESUME

ED 357 078 TM 019 857

AUTHOR Beasley, T. Mark; Leitner, Dennis W.
TITLE Nonparametric Test of Ordered Alternatives: Extension of Page's L Test for Two Groups of Unequal Size.
PUB DATE Apr 93
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Comparative Analysis; Equations (Mathematics); Estimation (Mathematics); Evaluators; Interrater Reliability; *Mathematical Models; *Nonparametric Statistics; *Sample Size; *Statistical Distributions
IDENTIFIERS A Priori Tests; *L Test; Two Group Test (Leitner Dayton); Type I Errors; *Unit Normal Approximation

ABSTRACT

The L statistic of E. B. Page (1963) tests the agreement of a single group of judges with an a priori ordering of alternative treatments. This paper extends the two group test of D. W. Leitner and C. M. Dayton (1976), an extension of the L test, to analyze difference in consensus between two unequally sized groups of judges. Exact critical values are tabulated for small numbers of treatments and judges, and a unit normal approximation is developed for larger samples. A brief computation example is also provided. Analysis of the two-tailed unit normal approximation shows that disparity between sample sizes often leads to underestimating the probability of a Type I error. However, as the number of treatments increases, the fit becomes better. This test, in comparison to two parametric competitors in a 2 by 4 mixed design, made fewer Type I errors when data were sampled from the normal, uniform, and exponential distribution, but it was also shown to be generally more conservative. The unit normal approximation is not severely affected by unequally sized groups having heterogeneous variances. The proposed test shows adequate power in conditions that do not favor parametric tests (i.e., interval level, normally distributed variables). Four tables present analysis results, and four figures illustrate the discussion. (SLD)

Nonparametric Test of Ordered Alternatives: Extension of Page's L Test for Two Groups of Unequal Size

T. Mark Beasley & Dennis W. Leitner

Southern Illinois University-Carbondale

Paper Presented at the Annual Meeting of the American Educational Research Association. April, 1993. Atlanta GA.

Abstract

Page's (1963) L statistic tests the agreement of a single group of judges with an *a priori* ordering of alternative treatments. This paper extends the two group test developed by Leitner & Dayton (1976) to analyze the difference in consensus between two unequally sized groups of judges. Exact critical values are tabled for small numbers of treatments and judges; a unit normal approximation (z_{LB})is developed for larger samples. A brief computation example couched in terms of an educational study is also provided. Analysis of the two-tailed unit normal approximation shows that disparity between sample sizes often leads to underestimating the probability of committing a Type I error; however, as the number of treatments increases the fit becomes better. This test in comparison to two parametric competitors in a 2 x 4 mixed design made fewer Type I errors when data were sampled from the normal, uniform, and exponential distribution, but it was also shown to be generally more conservative. Also in contrast to the parametric tests, the z_{LB} is not severely affected by the unequally sized groups having heterogeneous variances. Furthermore, the proposed test shows adequate power in conditions that do not favor parametric tests (i.e., interval level data; normally distributed variables).

Page's L test provides a nonparametric statistical index for the agreement of a single group of n rankings to a set of k alternatives (i.e., treatments, objects) with the following null hypothesis:

$$H_o: T_1 = T_2 \ldots = T_k \quad (1)$$

where T_i represents the mean for the ith treatment. The most general alternative hypothesis is that at least one pair among the k alternatives (T_i) is not equal. However, in many experiments an *a priori* ordered outcome of the k treatments can be expected from theoretical considerations and is thus of scientific interest. In such a case, the L test can be used and the alternative hypothesis can be phrased as such:

$$H_a: T_1 \leq T_2 \leq \ldots \leq T_k \quad (2)$$

with at least one strict inequality. Thus, the L test was specifically designed for use with data measured at or reduced to an ordinal level. This test of ordered hypotheses for multiple treatments is based on the linear contribution each ranking makes to the treatment sum of squares (Page, 1963). It has also been shown to be equivalent to the average Spearman rank-order correlation between the *a priori* ordering and the group of n rankings (Lyerly, 1952; Page, 1963). That is, the average of the n correlations between individual rankings and the *a priori* ordering is mathematically related to L. Furthermore, the L test has parametric analogs in experimental designs such as the randomized block design (Azzam, Awad, & Sarie, 1987; Hollander, 1967) and the repeated measures or split-plot design (Siegel & Castellan, 1988). Thus, the L test is easily computed, but few extensions to more complicated designs have been developed. Therefore, this procedure has been of questionable versatility, although it has been shown to be extremely robust because of its relationship to the Spearman coefficient (Page, 1963). Hollander (1967) showed that the asymptotic power of Page's L test compared to the t-test is .714 for k=3 and .955 as k approaches infinity.

In practice, Page's L test is used to statistically determine whether one group of n judges agrees with an *a priori* ordering of treatments. For example, this test could be used to examine whether four school board members (n=4) rank the importance of five

educational objectives (k=5) as defined by Bloom's Taxonomy of Educational Objectives (i.e., Bloom & Madaus, 1981). In Table 1, the hypothetical data are cast into two-way tables having n rows and k columns. Separately for each row (school board member), the 5 objectives are ranked. Each column of treatments (T_j) are summed and weighted by the order of the treatment. Then these weighted components are summed to form L which can be represented with the following formula:

$$L = \sum_{i=1}^{k} i \sum_{j=1}^{n} T_i \qquad (3).$$

Thus from the data in Table 1 (Panel A), L = 217 which is significant with $p < .001$ (see Page, 1963 for tables). Therefore, the null hypothesis stating that the mean ranking of the T_j's are equal across the k objectives was rejected and the alternative hypothesis stating that the school board members ranked at least one of the objectives in the *a priori* ordering was accepted.

If a second group, say concerned parents, were considered; two questions can be asked. First, to what degree does the consensus of the second group of judges (parents) agree with the *a priori* ranking of objectives? Second, do the two groups differ in the degree of their consensus with the ordered hypotheses. The first question can be answered by Page's L test applied to the second group of judges. Table 1 (Panel B) shows another hypothetical example using 4 parents ranking of the 5 objectives. The result (L = 185) was not significant at the .05 alpha level. Thus, there is not sufficient evidence to reject the null hypothesis that the 5 objectives were ranked equally.

The second question can be answered by an extension of Page's L, the L_D test (Leitner & Dayton, 1976) which is defined as the absolute value of the differences between L's for the two groups with the following null hypothesis:

$$\text{Ho: } L_1 - L_2 = 0 \text{ or Ho: } L_D = 0. \qquad (4)$$

The analysis for these two groups yields a results, $L_D = (217 - 185) = 32$, significant at the .05 alpha level (see Leitner & Dayton, 1976 for tables). Thus, the two groups (school board members and parents)

significantly differ in their degree of consensus to the *a priori* ordering of objectives as defined by Bloom's taxonomy. Furthermore, since school board members have a larger L statistic than do parents, they also have a significantly higher amount of agreement to the hypothesized ordering than do the parents. In certain situations, the L_D test of no differences in L statistics can be viewed as analogous to testing the linear trend interaction in a 2 by k mixed ANOVA. This test also assumes that L_D becomes large up to a maximum value when the groups disagree and that it is symmetric around zero. Thus, although Leitner and Dayton defined the L_D test as an absolute value, directional tests are possible. However, the symmetric conditions for the null hypothesis do not hold for unequal n's, which limits analyses to experiments with groups of equal sample size. In the example given, school board members, as opposed to concerned parents, are relatively rare and would have smaller sample sizes in most replications of this hypothetical study. Moreover, equal sample sizes are uncommon in most research studies. In the present example, if one parent with a ranking of (5,4,3,2,1) and thus an individual L equal to 37 was added, the resultant L would equal 220 and L_D would have an absolute value equal to 3 seemingly not of sufficient magnitude to be "significant" in comparison to the previous L_D. Yet, there appears to be quite a disparity between the priorities of the school board members and the parents although the L_D does not reflect this. To elucidate, if six school board members were to exactly agree with the theoretical *a priori* ordering of the 5 objectives, the resultant L would be 330. If eight concerned parents were also to exactly agree with these ordered hypotheses, the result would be an L = 440. Thus, L_D would equal 110 although every judge, regardless of background, agreed with the hypothesized ordering although there should be no differences detected. Thus, larger groups would have larger L's by virtue of size rather than linearity of ranks. One way to circumvent this problem is to scale the two L's to the same metric. In the present paper, it is proposed to use an "averaged L" to test differences with the following null hypothesis:

$$\text{Ho: } L_B = |\bar{L}_1 - \bar{L}_2| = 0 \qquad (5)$$

where, $\bar{L}_1 = \frac{L_1}{n_1}$ and $\bar{L}_2 = \frac{L_2}{n_2}$.

Page (1963) showed that the L has the following mean:

$$E(L) = \mu_L = \frac{nk(k+1)^2}{4} \qquad (6)$$

In scaling this L with a division by n the expected value for the "averaged L" is as follows:

$$E(\bar{L}_i) = \frac{k(k+1)}{4} \qquad (7).$$

Thus, the following can be elaborated:

$$E(L_B) = E(\bar{L}_1 - \bar{L}_2) = \frac{n_1k(k+1)^2}{4n_1} - \frac{n_2k(k+1)^2}{4n_2} = 0 \qquad (8).$$

From the third example, $L_B = 330/6 - 440/8 = 55 - 55 = 0$. Thus, there are no differences in the groups' rankings although their sample sizes differ. From the second example (Table 1, Panel C), $L_B = 217/4 - 220/5 = 10.25$, indicating that differences do indeed exist. More importantly, this formulation holds for both equal and unequal sample sizes and is symmetric around zero.

In addition, both Page (1963) and Leitner and Dayton (1976) developed unit normal approximations for larger samples. Page (1963) showed that the L test has the following variance:

$$\sigma^2(L) = \frac{nk^2(k+1)^2(k-1)}{144} \qquad (9).$$

Therefore, the unit normal approximation of the L test is the sample L minus the expected value of L (Eq. 1.6) and divided by the variance of L (Eq. 1.9) which reduces to:

$$z_L = \frac{12L - 3nk(k+1)^2}{k(k+1)\sqrt{n(k-1)}} \qquad (10)$$

For the Leitner and Dayton L_D test, the two separate Ls have the same expected value and therefore cancel out. Thus, the absolute difference between the L's is divided by a denominator based on the standard deviation of the differences in L's. Assuming the two groups to be independent, L_D should have a variance equal to $\sigma^2(L_1) + \sigma^2(L_2)$ which, with equal number of judges in each group, becomes

two times the variance defined in equation 12. Thus the L_D can be formulated as:

$$z_{LD} = \frac{|L_1 - L_2|}{k(k+1)} \sqrt{\frac{72}{n(k-1)}} \qquad (11).$$

As previously mentioned, the development of an analogous test for unequal sample sizes the difference between L's was redefined in terms of L_1/n_1 and L_2/n_2. Thus, L_B has the following variance:

$$\sigma^2(L_B) = \sigma^2(\bar{L}_1 - \bar{L}_2) = \sigma^2(\bar{L}_1) + \sigma^2(\bar{L}_2) \qquad (12).$$

Using equation 1.9 and the property of multiplying a constant to the variance of a set yields:

$$\sigma^2(\bar{L}_i) = \frac{n_i k^2(k+1)^2(k-1)}{144\, n_i^2} = \frac{k^2(k+1)^2(k-1)}{144\, n_i} \qquad (13).$$

Thus for two unequally sized groups, $\sigma^2(L_B)$ reduces to:

$$\sigma^2(L_B) = \sigma^2(\bar{L}_1 - \bar{L}_2) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left[\frac{k^2(k+1)^2(k-1)}{144}\right] \qquad (14)$$

and the unit normal approximation is:

$$Z_{LB} = \frac{12\,|\bar{L}_1 - \bar{L}_2|}{k(k+1)\sqrt{(k-1)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (15)$$

which is equivalent to the Leitner and Dayton test (Eq. 1.11) when the sample sizes are equal. In the present example (Table 1 including Panel C), the result, $z_{LB} = 3.06$, is statistically significant at the .05 alpha level, assuming a unit normal approximation.

Tests Related to L

Since Page's L test an ordered hypothesis of monotonic trends in data, it also tests the linear trend in ranks. Given linear effects, the L test is also analogous to regression analyses. Since most rank tests are considered to be alternatives to parametric procedures based on the General Linear Model, there are several nonparametric tests related to Page's L test.

Single Group Tests. Similar to the k by n layout analyzed by Page's L, the Friedman two-way ANOVA by ranks (Friedman, 1937) tests a null hypothesis that the k matched samples (or repeated measures) have been drawn from the same population. The

alternative hypothesis, however, is more general than that of the L test and states that at least one pair among the k repeated measures has different medians. The Friedman statistic is based on Kendall's coefficient of concordance (Kendall & Babbington-Smith, 1937) which expresses the degree of association among k sets of rankings for n judges. The Friedman test is formulated as such:

$$X^2 = \left[\frac{12 \sum_{j=1}^{k} \sum_{i=1}^{n} (T_{ij})^2}{n\,k\,(k+1)} \right] - 3n(k+1) \qquad (16),$$

where T_{ij} is the individual rank given from 1 to k for ith judges. This statistic approximates a chi-squared distribution with k-1 degrees of freedom and differs from Page's L in that the column totals are squared rather than multiplied by some *a priori* coefficient ranging from 1 to k (see Eq. 1.3 for comparison). Thus the Friedman X_F^2, used for testing the significance of the differences among treatments, is closely related to an omnibus F-tests of the treatment (repeated measures) sum of squares (SS_T) in the parametric split plot design. Indeed, it can be shown that X_F^2 is equivalent to:

$$X_F^2 = \frac{12\,SS_T}{k\,(k+1)} \qquad (17).$$

Similarly, the square of the L test unit normal approximation (Eq. 1.10 squared) can be proved equivalent to:

$$z_L^2 = X_L^2 = \frac{12}{k(k+1)} \text{ x (linear contribution of } SS_T) \qquad (18).$$

Given this restriction of linearity across the k alternatives, the L test is related to correlation and regression (Page, 1963). Lyerly (1952) proposed a statistic called average rho, $\bar{r}$, which was equivalent to the average Spearman rank-order correlation in a set of n judges. The equivalence of Page's L to the Lyerly and Spearman statistics can be demonstrated. Page (1963) showed that L was equivalent to $\bar{r}$:

$$\bar{r} = \frac{12L}{nk(k^2-1)} - \frac{3(k+1)}{(k-1)} \qquad (19)$$

One variant of Page's L test has been proposed in order to better separate the rankings in middle of the distribution (Azzam et al.,

1987). The B statistic was proposed because certain rankings which are distinct from each other receive the same individual L. For example with k=4, a ranking of [1, 3, 4, 2] and a ranking of [2, 3, 1, 4] both receive an individual L of 27. Thus, although the rating process that underlies these rankings may be quite different, Page's L does not distinguish between the two. Therefore, squaring the *a priori* ordering coefficient was proposed such that:

$$B = \sum_{i=1}^{k} i^2 \sum_{j=1}^{n} T_i \qquad (20)$$

This test has been shown to be generally as powerful as Page's L (Azzam et al., 1987), but it has a rather complicated unit normal approximation. Furthermore, the B statistic involves squaring the ordering coefficient; thus, agreement to the latter alternatives is given more weight and is interpreted as more important than agreement to the first alternatives. Because of the recency of this test along with the implicit weightings and differentiations that are made, its efficacy in practical situations has not been assessed.

Hollander (1967) proposed a rank procedure for testing ordered alternatives which has since been classified as an A-type test because the rankings are taken among blocks. This differs from Page's L, a W-type test, in which the rankings are taken within a block (Pirie, 1974). This A-type procedure involves ranking the differences among blocks then summing the ranks to form the Y statistic. Hollander's Y is not distribution free, but it is asymptotically normal with the following expected value and variance:

$$\mu(Y) = \frac{k(k-1)n(n+1)}{12} \qquad (21)$$

$$\sigma^2(Y) = \frac{n(n+1)(2n+1)(3k-2)\rho_0^n(F)}{144} \qquad (22)$$

where $\rho_0^n(F)$ is a factor derived from the sampled distribution, F. The values for this factor is reported in Hollander (1967). The asymptotic efficiency of Y relative to the t-test is greater than .864 for all distribution functions, F, and all numbers of alternatives (k). When F is normal, the Asymptotic Relative Efficiency (ARE) with

respect to the t-test range from .963 to an upper limit of .989 as k approaches infinity. Therefore, Y outperforms Page's L under these conditions (Hollander, 1967). However, when F is a uniform or exponential distribution function, ARE values are at least as likely to favor Page's L as Hollander's Y. In fact, which test performs better apparently depends on the sampled distribution and the values of k and n. In any case, the differences in power cannot be expected to be great. Thus, Page's L and W-type tests, in general, are favored over Hollander's Y (and other A-type tests) specifically because W-type tests are: a). distribution free; b). more able to control Type I error rates; and c). easily computed (Pirie, 1974).

Multiple Group Tests. Jonkheere (1954) proposed a test of ordered alternatives for k independent samples. This statistic tests the null hypothesis that the medians are the same across groups (samples). The alternative hypotheses is that the medians are ordered in magnitude with at least one strict inequality. This procedure uses a Mann-Whitney count method and is based on the average Kendall rank-order correlation (Kendall's tau) between the observed ranking of the i^{th} judge and the *a priori* ordering. Hollander (1967) showed that the asymptotic power of Jonkheere's test as compared to the t-test is similar to that of Page's L with an ARE of .694 for k=3 and as an ARE of .955 as k approaches infinity. Furthermore, since Jonkheere's test is based on Kendall's rank-order coefficient, Page's L can be shown to be more powerful because of the L's within-subject design. As compared to Kendall's tau, Page's L has an ARE equal to:

$$ARE(L|\tau) = \frac{k(2k+5)}{2(k+1)^2}$$

which reaches its maximum at k=5 and never falls below zero.

The Schucany and Frawley (1973) model is based on Page's L and is designed to test differential concordance in terms of the correlation between ranks of k alternatives assigned by two independent groups of judges. The statistic takes the product of the two separate rankings totaled over judges and then sums over the k alternatives:

$$S = \sum_{j=1}^{k} R_{1j}R_{2j} \qquad (23).$$

A unit normal approximation was formulated as such:

$$z_S = \frac{12S - 3n_1n_2k(k+1)^2}{k(k+1)\sqrt{n_1n_2(k-1)}} \qquad (24).$$

Unlike the test presently proposed which has a null hypothesis of no between group differences in L, the Schucany and Frawley model has been criticized because it tests the null hypothesis that there is no concordance (Serlin & Marascuilo, 1983). Thus, a between group concordance of disconcordance with the *a priori* ordering is a possible result. Serlin and Marascuilo (1983) point out, "It is hard to conceive of concordance between groups when there is no evidence that there is concordance within groups" (p. 194).

Hollander and Sethuraman (1978) have also questioned the Schucany and Frawley model because it tests for positive rank-order correlation in the mean ranks as an alternative. As an alternative, Hollander and Sethuraman provided a two group procedure which tests the identity of mean ranks across the k alternatives as the null hypothesis. Serlin and Marascuilo (1983) extended this method for multi-group situations and also developed planned and post-hoc comparison procedures. These multiple comparison procedures are capable of testing group differences at each of the k levels of treatment alternatives and are thus analogous to simple main effects in the ANOVA. Although the computation of these comparison procedures are relatively simple, the omnibus test for the Hollander and Sethuraman as well as the Serlin and Marascuilo formulations are based on multivariate procedures and are rather complex.

Since Page's L is a test of the linearity of ranks, power analyses will proceed by generating rankings around different linear and monotonic effects. Then a comparison of the L_B tests to the mixed design ANOVA with linear and "double-ends" monotonic (Gaito, 1965) interaction contrasts will be completed. The relative power of the L_B tests are not expected to exceed the parametric tests in many cases because within ranking and within-group variances will

initially be kept homogeneous. However, the well-known effects of violating the homogeneity of variance assumption in combination with unequal sample sizes (Glass, Peckham, & Sanders, 1972) could become an issue. This is possible since nonparametric analyses have shown to be more powerful (or robust) under such violations (Boneau, 1962).

Methods

The distributions of L were generated for three alternatives (k=3) from a sample size of 2 (n=2) to as large as a sample size of n=10. For k=4, the generated distributions of L ranged from n=2 to n=8 and for k=5, from n=2 to n=6. For the two group situation all possible pairwise differences in L's (L_B distributions) were generated within a given number of alternatives. The critical values at the .10, .05, and .01 alpha levels were determined by finding the point in the L_B distribution that did not exceed the 90th, 95th or 99th percentile, respectively. These critical values were tabled along with the p-value derived from the z_{LB} to show the approximation of this statistic. To demonstrate the fit of the two-tailed unit normal approximation (z_{LB}), the ogive of the actual L_B distributions could be compared the ogive of theoretical distribution it is supposed to approximate (i.e., $X^2_{(1)}$) and subsequently the Kolmogorov-Smirnov (K-S) test could be used to test the fit of these statistic. However, it is important to note that the sample size for a given distribution is $k!^N$, where N is the total sample size, so that the K-S test will be highly sensitive to minor deviations at any point in the distribution. Furthermore, in using the K-S test, fit as a null hypothesis can only be falsified and thus retention of such a null hypothesis does not prove the approximation of the statistic. Moreover, since significance testing in general uses cumulative proportions of theoretical or actual distributions to establish critical regions for the rejection of null hypotheses, the fit at the upper end of the ogives is of more practical interest. Therefore, to examine the fit of these statistics at the upper end in context with the disparity of sample sizes, the difference between the actual cumulative proportion above the L_B critical value and the p-value of the approximate statistic (z_{LB}) at the .10, .05, and

.01 alpha levels are plotted as a function of the ratio of the largest to total sample size.

To examine the effects of violating the parametric normality assumption on the Type I error rate, data for 4 alternatives with no difference between the L's of two equally sized groups of n=8 (no trend interaction for the ANOVA) were randomly generated from a normal, uniform and exponential distribution and replicated 1,000 times. The rejection rates of the z_{LB} and the ANOVA are compared at the .041 alpha level since this is the actual proportin above the critical value for k=4 and equal n's of 8 (see Table 3). The cell parameters were [2, 3, 4, 1] for both groups so that there is no expected interaction and both groups have the same expected value for L, E(L)=24. In sampling from the normal distribution, data were generated around these cell parameters with equal cell variances, $\sigma^2_{(wc)}=4$. Using the same parameter seeds for generating the uniformly and exponentially distributed data, the expected values of the cell parameters change only by a constant (i.e., E(x) + 1/2 for the uniform distribution and E(x) + $\sigma_{(wc)}$ for the exponential distribution) so that the expected ranks, E(L), and the expected differences in means are not changed. In using the exponential random generator the variance is not affected, but the distribution becomes positively skewed, while the variance for the uniform distribution is changed so that the within cell variance is $\sigma^2_{(wc)}/12$ (i.e., 3 in this case) for the uniform distribution (Freund & Walpole, 1987).

To examine the effects of heterogeneity of variance with unequal sample sizes on the Type I error rate at the three levels of alternatives, normally distributed data with no differences in the 2 L's (same as the previous parameters) were randomly generated at differing ratios of variances and sample sizes. With the parameter average within-cell variance held constant at 9, the within-cell variance ratio of Group 1 to Group 2 are examined at .2 (3/15), .33 (4.5/13.5), 1 (9/9), 3 (13.5/4,5), and 5 (15/3). With total sample size held constant at 20, largest (Group 1) to total sample size ratios are examined at .5 (n_1 =10; n_2=10), .7 (n_1 =14; n_2=6), .8 (n_1 =16; n_2=4), and .9 (n_1 =18; n_2=2). All levels of the sample size and variance ratios were crossed and 1,000 replications in each cell were performed for

both the z_{LB} and the ANOVA. For the .05 alpha level, deviations from 50 (5%) rejections will indicate the "conservative" or "liberal" nature resultant from the violations of this assumption.

To analyze power, the two group test for 4 alternatives was compared to linear polynomial and "double-ends" monotonic (Gaito, 1965) interaction contrasts with a 2 x 4 mixed ANOVA. Two basic situations are presented. In one case cell means were randomly generated around the following equally spaced parameters which is analogous to having interval level data:

Group 1	[1, 3, 4, 2]	E(L)=27	Equally Spaced
Group 2	[3, ,2, 4, 1]	E(L)=23	Parameters

In this case the rankings of the data and the original data itself would yield the same results in any parametric procedure since with equally spaced parameters rankings are simply a linear transformation of the data. But in ranking a set of alternatives, the process underlying this ranking procedure may not be based on equally spaced parameters or the data simply may not be measured at an interval level. Therefore, data with the same expected values of L were generated around the following unequally spaced parameters:

Group 1	[5, 10, 12, 7]	E(L)=27	Unequally Spaced
Group 2	[10, 9, 21, 8]	E(L)=23	Parameters

Thus, the expected value of differences in L's are same in both case, $E(L_B)=4$. In both cases, the distributions are sampled from the normal, uniform, and exponential distributions, while the respective cell parameters are held constant. The effects of increased sample size on power are examine by using n=4, n=8, n=16, and n=32. In the analysis of these effects when the data are sampled from a normal distribution, $\sigma^2_{(WC)}$ was held constant at 7.51 for the equally spaced parameters and at 1.55 for the unequally spaced parameters. These two values were used because a $\sigma^2_{(WC)}=7.55$ gives the linear interaction contrast for the equally spaced parameters with an n=4 a

Non-Centrality Parameter ,NCP) equal to the test's .05 critical value, NCP(1,18)=4.41. A $\sigma^2_{(wc)}$ =1.55 does the same for the monotonic interaction contrast of the unequally spaced parameters. In keeping the same generation seeds, the uniformly distributed equally spaced parameters have $\sigma^2_{(wc)}$=.625, while for the unequally spaced parameters $\sigma^2_{(wc)}$=.129.

Results

Tables 2, 3, and 4 show the exact critical values, the actual cumulative proportion above the critical value, and the p-values of the unit normal approximate at nominal alpha levels of ,10, .05, and .01 for 3, 4, and 5 alternatives, respectively. To examine the fit of the large sample approximate (z_{LB}), the difference betweer. the actual cumulative proportion above the critical value and the p-value of the approximate statistic at the .10, .05, .01 alpha levels are plotted as a function of the ratio of the largest to total sample size. Figure 1 (first panel) shows that for 3 alternatives the values are all negative ranging from about -.02 to 0 at the .10 alpha level which is reflected by the theoretical distribution having higher ordinates than the actual distributions. At the .05 level, the fit becomes better with both positive and negative values basically centered around zero. At the .01 level, the values are positive ranging from 0 to .01 which is reflected by the actual distribution going above the theoretical $X^2_{(1)}$. The same basic results can be seen for 4 and 5 alternatives (Fig. 1; second and third panels), but the differences between the theoretical and actual cumulative proportions are smaller, demonstrating that the approximation is closer as the number of alternatives increases. Another interesting effect is the linear relationship between these distributional differences and sample size ratio most notable at .05 and .01 alpha level for k=3. Although the approximation of the distribution can be used to describe these effects, it can also be explained in terms of the well-known heterogeneity of variance effects. That is, in the formula for within-cell variance (Eq. 13) larger samples by definition have smaller variances. Also since the distributional differences were calculated as the actual cumulative

proportion minus the theoretical cumulative proportion, positive values mean that the p-value calculated from z_{LB} is less than the actual cumulative proportion below that particular difference in mean L's which in practice would lead to more rejections. Therefore, positive values can be viewed as indicating more "liberal" tests while negative values reflect more "conservative" rates of rejection. Thus the positive relationship between the distributional differences and sample size ratio at these particular alpha levels indicates that as the disparity between group sample sizes increase and the larger sample by definition has less within cell variance, the test becomes more "liberal" which is consistent with the effects of heterogeneity of variance.

For a 2 x 4 design (collapsed to the L_B test for 4 alternatives), Figure 2 shows the proportions of rejections under null conditions for L_B and for linear and monotonic interactions contrasts when the data are sampled from the normal, uniform, and exponential distributions. Across all conditions, L_B commits fewer Type I errors than either of the parametric tests, but it also makes fewer than the expected number of false positives. Thus, the test is somewhat "conservative". When sampled from the normal distribution and especially the uniform distribution, the results reflect the conservative nature of L_B. By contrast the three different tests show similar results when the data are sampled from the exponential distribution.

Figure 3 replicates the effects of heterogeneity of variance in combination with disparate sample sizes and shows that z_{LB} is robust to such violations. In Figure 3, the horizontal dashed line is at the alpha level of .05. Major deviations from this line show the effects of unequal samples sizes and variances. The vertical dashed line in each panel is at 1.0 where the within-cell variances are equal for the two groups and unequal cell frequencies should have minor effects. In the upper left panel of Figure 3, the largest to total sample size ratio is .5, the situation in which the sample sizes are equal. At this point all three tests keep the nominal alpha level of .05, but as expected, as the largest to total sample size ratio increases heterogeneity of variance affects the rejection rates of the

parametric tests. For the parametric tests, when the larger sample has the smaller variance, the tests are more liberal and the Type I error rate increases to as high as 33% (Fig. 3, lower right panel). When the large sample has the smaller variance the rejection rates of the parametric tests approach zero. For, z_{LB}, however, the rejection rates range from .075 to .01 which given the conservative nature of this test falls within sampling error. Thus, the proposed test are robust to the violation of the heterogeneity of variance assumption in normally distributed data.

Figure 4 shows the power of z_{LB} relative to linear and monotonic interaction contrasts under various conditions. As expected, since equally spaced parameters are analogous to having interval level data, the parametric tests dominate, especially when the data are sample from the normal distribution. However, given a moderate sample (cell sizes of n = 16) drawn from either the uniform or exponential distribution, the z_{LB} has comparable power. When parameters are unequally spaced, linear contrasts cannot compete for two reasons. If the data are not interval level, linear test are not designed to readily detect their effects;however, if the data are interval level, but the parameters are not equally spaced, the use of linear test is a misspecification and therefore inappropriate. Thus, the monotonic interaction contrasts are definitely more powerful in these situations. Under the conditions that favor parametric tests (i.e., normally distributed data), the monotonic contrasts show more power, although z_{LB} becomes comparable around a sample size of n=32. When data are sample form the exponential distribution, z_{LB} and the test of monotonic interaction are of comparable power at n=16. When data under these simulated conditions are sample from the uniform distribution, both tests are very powerful.

Discussion

These results demonstrate that the proposed procedure for testing differences in *a priori* monotonic trends or linear trends of ranks by testing the differences in Page's L approximates the commonly used unit normal and chi-square distributions, is robust to the violations most deleterious to parametric tests (i.e., heterogeneity of variance), and can be rather powerful under certain distributional

conditions. Furthermore, it makes less restrictive assumptions about the shape of the sampled distribution and is easy to compute. The results showed that as the number of treatments increase the fit of the unit normal approximation (z_{LB}) becomes better, but even for three alternatives the p-value from z_{LB} is no more than 2% different than the proportion above the actual critical value. The performance of this test in the 4 alternative situation was compared to a 2 x 4 mixed design and was shown to make fewer Type I errors than its parametric competitors when data were randomly generated from the normal, uniform, and exponential distributions. This nonparametric test, however, was also shown to be generally more conservative, especially with uniformly distributed variables. In contrast to the parametric tests, z_{LB} is not seemingly affected by unequally sized samples having unequal variances. Furthermore, z_{LB} shows adequate power in conditions that do not favor parametric tests (i.e., interval level data; normally distributed variables). Specifically, when data are sampled from either the uniform of exponential distribution, z_{LB} is of comparable power with as few as 16 judges per group.

This extension should prove beneficial in educational research for many reasons. First, well-planned educational research often involves directional and *a priori* hypotheses about multiple treatment effects. Secondly, applied educational research often involves smaller samples which are unlikely to be equal in size. Also, the dependent variables used to capture educational and psychological phenomena mostly involve ordinal scale measurement which with moderate sample sizes are not technically amenable to most parametric procedures.

References

Azzam, A. Awad, A., & Sarie, T. (1987). A non-parametric test for ordered alternatives in a randomized block design. Dirasat, 15, 39-47.

Bloom, B. S., & Madaus, G. F. (1981). Evaluation of student learning. New York: McGraw-Hill.

Boneau, C. A. (1962). A comparisonof the power of the *U* and *t* test. Psychological Review, 69, 246-256.

Freund, J. E., & Walpole, R E. (1987). Mathematical statistics (4th Ed.). Englewood Cliffs, NJ: Prentice-Hall.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32, 675-701.

Gaito, J. (1965). Unequal interval and unequal n in trend analysis. Psychological Bullentin, 63, 125-127.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the analysis of variance and covariance. Review of Educational Research, 42, 237-288.

Hollander, M. (1967) Rank tests for randomized blocks when the alternatives have an *a priori* ordering. Annals of Mathematical Statistics, 38, 867-877

Hollander, M., & Sethuraman, J. (1978). Testing for agreement between two groups of judges. Biometrika, 65, 403-411.

Jonkheere, A. R. (1954). A test of significance for the relation between *m* rankings and *k* ranked catregories. British Journal of Statistical Psychology, 7, 93-100.

Kendall, M., & Babbington-Smith, B. (1939). The problem of *m* rankings. Annals of Mathematical Statistics, 10, 275-287.

Leitner, D. W., & Dayton, C. M. (April, 1976). A two-group test of ordered hypotheses for multiple treatments: An extension of Page's L test. Paper presented at the meeting of the American Educational Research Association. San Francisco, CA.

Lyerly, S. B. (1952). The average Spearman rank correlation coefficient. Psychometrika, 17, 412-428.

Page, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. Journal of American Statistical Association, 58, 216-230.

Pirie, W. R. (1974). Comparing rank tests for ordered alternatives in randomized blocks. Annals of Statistics, 2, 374-382.

Schucany, W. R., & Frawley, W. H. (1973). A rank test for two group concordance. Psychometrika, 38, 249-258.

Serlin, R. C., & Marascuilo, L. A. (1983). Planned and post-hoc comparisons in tests of concordance and disconcordance for G groups of judges. Journal of Educational Statistics, 8, 187-206.

Siegel, S., & Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences (2nd ed.). New York: McGraw-Hill.

Table 1. Hypothetical data for tests of ordered alternatives.

Panel A. Hypothetical school board data

School Board Member	Objectives (k= 5) T_1	T_2	T_3	T_4	T_5
S_{11}	1	3	2	4	5
S_{12}	1	2	3	5	4
S_{13}	2	1	3	4	5
S_{14}	1	2	3	4	5
$n_1=4$	$\Sigma T_1=5$	$\Sigma T_2=8$	$\Sigma T_3=11$	$\Sigma T_4=17$	$\Sigma T_5=19$

$L_1 = 217 = (1)(5) + (2)(8) + (3)(11) + (4)(17) + (5)(19)$

Panel B. Hypothetical parent data

Parent	Objectives (k= 5) T_1	T_2	T_3	T_4	T_5
S_{21}	2	1	3	4	5
S_{22}	3	4	5	1	2
S_{23}	1	5	4	3	2
S_{24}	3	2	5	1	4
$n_2=4$	$\Sigma T_1=9$	$\Sigma T_2=12$	$\Sigma T_3=17$	$\Sigma T_4=9$	$\Sigma T_5=13$

$L_2 = 185 = (1)(9) + (2)(12) + (3)(17) + (4)(9) + (5)(13)$

$L_D = (217 - 185) = 32$

From equation 11, $z_{LD} = \frac{32}{5(5+1)}\sqrt{\frac{72}{4(5-1)}} = 2.26$

Panel C. Data for fifth parent.

S_{25}	5	4	3	2	1

$L_2 = 220 = (1)(14) + (2)(16) + (3)(20) + (4)(11) + (5)(14)$

$L_B = 217/4 - 220/5 = 10.25$

From equation 15, $z_{LB} = \frac{12(10.25)}{5(5+1)\sqrt{(5-1)(\frac{1}{4}+\frac{1}{5})}} = 3.06$

Table 2. Critical Values, Actual Proportions of the Distribution, and Alpha from the Unit Normal Approximation for k= 3

k=3

		n_1 = 3			n_1 = 4			n_1 = 5			n_1 = 6		
n_2	Nominal Alpha	CV	% Dist.	Approx. Alpha	CV	% Dist.	Approx. Alpha	CV	% Dist.	Approx. Alpha	CV	% Dist.	Approx. Alpha
2	.10	2.34	.0792	.0707	2.25	.0823	.0662	2.00	.0986	.0910	2.00	.0966	.0833
	.05	2.67	.0422	.0389	2.50	.0470	.0412	2.40	.0396	.0425	2.34	.0438	.0433
	.01	3.34	.0059	.0142	3.25	.0047	.0080	2.90	.0085	.0142	2.84	.0085	.0139
3	.10	2.34	.0576	.0433	1.84	.0998	.0896	1.80	.0858	.0814	1.84	.0795	.0668
	.05	2.67	.0262	.0209	2.17	.0470	.0449	2.07	.0478	.0454	2.17	.0337	.0303
	.01	3.34	.0031	.0039	2.75	.0087	.0109	2.60	.0093	.0118	2.67	.0063	.0077
4	.10				2.00	.0590	.0455	1.60	.0966	.0917	1.59	.0914	.0828
	.05				2.25	.0309	.0244	1.90	.0468	.0452	1.84	.0478	.0446
	.01				2.75	.0062	.0060	2.40	.0098	.0114	2.34	.0097	.0106
5	.10							1.60	.0933	.0736	1.44	.0988	.0942
	.05							2.00	.0314	.0253	1.70	.0478	.0471
	.01							2.40	.0081	.0073	2.20	.0091	.0102
6	.10										1.50	.0822	.0662
	.05										1.83	.0302	.0247
	.01										2.17	.0090	.0080

		n_1 = 7			n_1 = 8			n_1 = 9			n_1 = 10		
2	.10	1.93	.0974	.0879	2.00	.0805	.0736	1.89	.0903	.0875	1.90	.0902	.0828
	.05	2.22	.0476	.0508	2.25	.0406	.0442	2.12	.0498	.0562	2.20	.0379	.0446
	.01	2.72	.0088	.0167	2.75	.0064	.0139	2.62	.0080	.0182	2.60	.0080	.0176
3	.10	1.67	.0928	.0877	1.63	.0937	.0896	1.67	.0879	.0771	1.57	.0971	.0924
	.05	1.95	.0453	.0454	1.88	.0493	.0502	1.89	.0489	.0451	1.84	.0489	.0489
	.01	2.43	.0093	.0128	2.38	.0094	.0131	2.45	.0068	.0095	2.30	.0088	.0135
4	.10	1.50	.0946	.0906	1.50	.0973	.0833	1.42	.1001	.0955	1.45	.0888	.0830
	.05	1.75	.0492	.0484	1.75	.0494	.0433	1.67	.0500	.0499	1.65	.0502	.0486
	.01	2.25	.0092	.0111	2.25	.0090	.0094	2.14	.0099	.0188	2.10	.0103	.0121
5	.10	1.40	.0943	.0909	1.35	.0978	.0940	1.31	.1000	.0965	1.40	.0810	.0707
	.05	1.63	.0500	.0492	1.60	.0476	.0472	1.56	.0493	.0486	1.60	.0437	.0389
	.01	2.10	.0096	.0107	2.05	.0096	.0110	2.00	.0097	.0112	2.00	.0098	.0098
6	.10	1.31	.0990	.0960	1.29	.0963	.0908	1.28	.0938	.0865	1.23	.0959	.0913
	.05	1.55	.0501	.0492	1.54	.0453	.0435	1.50	.0471	.0442	1.47	.0460	.0446
	.01	2.00	.0102	.0110	1.96	.0097	.0103	1.94	.0087	.0091	1.87	.0098	.0106
7	.10	1.43	.0719	.0588	1.21	.1001	.0971	1.19	.0976	.0948	1.16	.0998	.0968
	.05	1.57	.0459	.0376	1.45	.0493	.0481	1.40	.0501	.0500	1.37	.0493	.0491
	.01	2.00	.0093	.0082	1.88	.0097	.0104	1.81	.0102	.0103	1.78	.0100	.0110
8	.10				1.25	.0929	.0771	1.15	.0969	.0934	1.13	.0975	.0935
	.05				1.50	.0408	.0339	1.35	.0497	.0499	1.33	.0497	.0482
	.01				1.88	.0091	.0080	1.76	.0097	.0103	1.73	.0096	.0101
9	.10							1.22	.0797	.0668	1.08	.0980	.0971
	.05							1.44	.0361	.0303	1.28	.0493	.0492
	.01							1.78	.0087	.0077	1.67	.0099	.0103
10	.10										1.10	.0968	.0820
	.05										1.30	.0472	.0398
	.01										1.70	.0082	.0072

Table 3. Critical Values, Actual Proportions of the Distribution, and Alpha from the Unit Normal Approximation for k= 4

k=4

		n1 = 3			n1 = 4			n1 = 5		
n2	Nominal Alpha	CV	% Dist.	Approx. Alpha	CV	% Dist.	Approx. Alpha	CV	% Dist.	Approx. Alpha
2	.10	4.50	.0935	.0876	4.25	.0998	.0887	4.10	.0936	.0894
	.05	5.33	.0437	.0431	5.00	.0479	.0455	4.80	.0455	.0469
	.01	6.67	.0087	.0112	6.25	.0097	.0122	6.00	.0088	.0126
3	.10	4.33	.0765	.0660	3.75	.0928	.0887	3.53	.0975	.0931
	.05	5.00	.0380	.0339	4.33	.0501	.0498	4.14	.0502	.0499
	.01	6.33	.0065	.0072	5.58	.0096	.0113	5.33	.0095	.0114
4	.10				3.50	.0986	.0865	3.25	.0963	.0933
	.05				4.25	.0416	.0367	3.80	.0501	.0497
	.01				5.50	.0067	.0071	4.95	.0093	.0106
5	.10							3.20	.0912	.0797
	.05							3.80	.0414	.0374
	.01							4.80	.0086	.0085

		n1 = 6			n1 = 7			n1 = 8		
2	.10	4.00	.0975	.0897	3.93	.0928	.0896	3.88	.0958	.0895
	.05	4.67	.0473	.0477	4.50	.0491	.0519	4.50	.0467	.0486
	.01	5.83	.0091	.0133	5.64	.0092	.0148	5.63	.0082	.0137
3	.10	3.50	.0946	.0864	3.33	.0974	.0943	3.25	.0992	.0963
	.05	4.17	.0433	.0412	3.90	.0495	.0500	3.83	.0490	.0498
	.01	5.17	.0101	.0114	5.00	.0096	.0121	4.88	.0099	.0126
4	.10	3.17	.0939	.0892	3.00	.0999	.0973	3.00	.0970	.0897
	.05	3.75	.0452	.0442	3.57	.0483	.0484	3.63	.0423	.0403
	.01	4.75	.0098	.0108	4.57	.0101	.0115	4.63	.0081	.0089
5	.10	2.90	.0995	.0971	2.80	.0998	.0976	2.73	.0998	.0978
	.05	3.43	.0497	.0495	3.31	.0500	.0499	3.23	.0500	.0500
	.01	4.47	.0096	.0106	4.31	.0096	.0107	4.20	.0096	.0107
6	.10	2.83	.0992	.0891	2.67	.0987	.0968	2.63	.0951	.0922
	.05	3.10	.0393	.0357	3.17	.0487	.0486	3.08	.0487	.0478
	.01	4.33	.0096	.0093	4.10	.0099	.0108	4.00	.0096	.0103
7	.10				2.71	.0867	.0786	2.48	.0981	.0966
	.05				3.14	.0457	.0417	2.93	.0500	.0500
	.01				4.00	.0099	.0095	3.80	.0101	.0109
8	.10							2.50	.0913	.0833
	.05							3.00	.0410	.0377
	.01							3.75	.0097	.0094

Table 4. Critical Values, Actual Proportions of the Distribution, and Alpha from the Unit Normal Approximation for k= 5

k=5

		n_1 = 3			n_1 = 4			n_1 = 5			n_1 = 6		
n_2	Nominal Alpha	CV	% Dist.	Approx. Alpha	CV	% Dist.	Approx. Alpha	CV	% Dist.	Approx. Alpha	CV	% Dist.	Approx. Alpha
2	.10	7.67	.0975	.0930	7.50	.0886	.0833	7.00	.0991	.0943	6.83	.0990	.0942
	.05	9.00	.0487	.0486	8.75	.0437	.0433	8.20	.0498	.0500	8.00	.0500	.0500
	.01	11.50	.0093	.0118	11.00	.0087	.0111	10.40	.0098	.0129	10.17	.0093	.0128
3	.10	7.00	.0945	.0864	6.33	.1001	.0972	6.13	.0966	.0930	6.00	.0954	.0897
	.05	8.33	.0439	.0412	7.50	.0496	.0495	7.20	.0493	.0486	7.00	.0492	.0477
	.01	10.67	.0082	.0090	9.75	.0099	.0114	9.20	.0098	.0112	9.00	.0095	.0109
4	.10				6.00	.0969	.0897	5.60	.0989	.0950	5.25	.0970	.0933
	.05				7.25	.0427	.0403	6.60	.0501	.0491	6.33	.0500	.0497
	.01				9.25	.0084	.0089	8.55	.0100	.0108	8.25	.0095	.0106
5	.10							5.40	.0984	.0877	5.03	.0981	.0964
	.05							6.37	.0478	.0430	5.93	.0499	.0488
	.01							8.20	.0097	.0095	7.73	.0098	.0106
6	.10										4.83	.0964	.0941
	.05										5.67	.0500	.0497
	.01										7.50	.0089	.0094

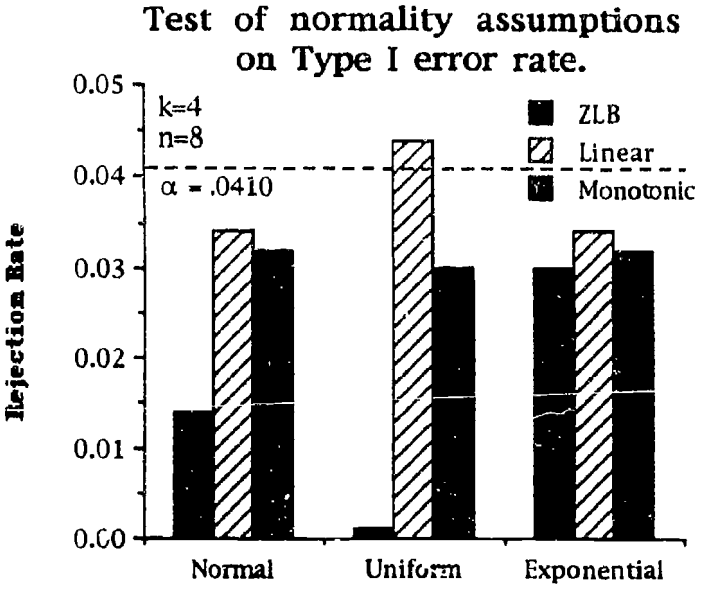

Figure 1. Null hypothesis rejection rates under normal, uniform, and exponential sampling conditions for L_B, linear, and monotonic interaction contrasts.

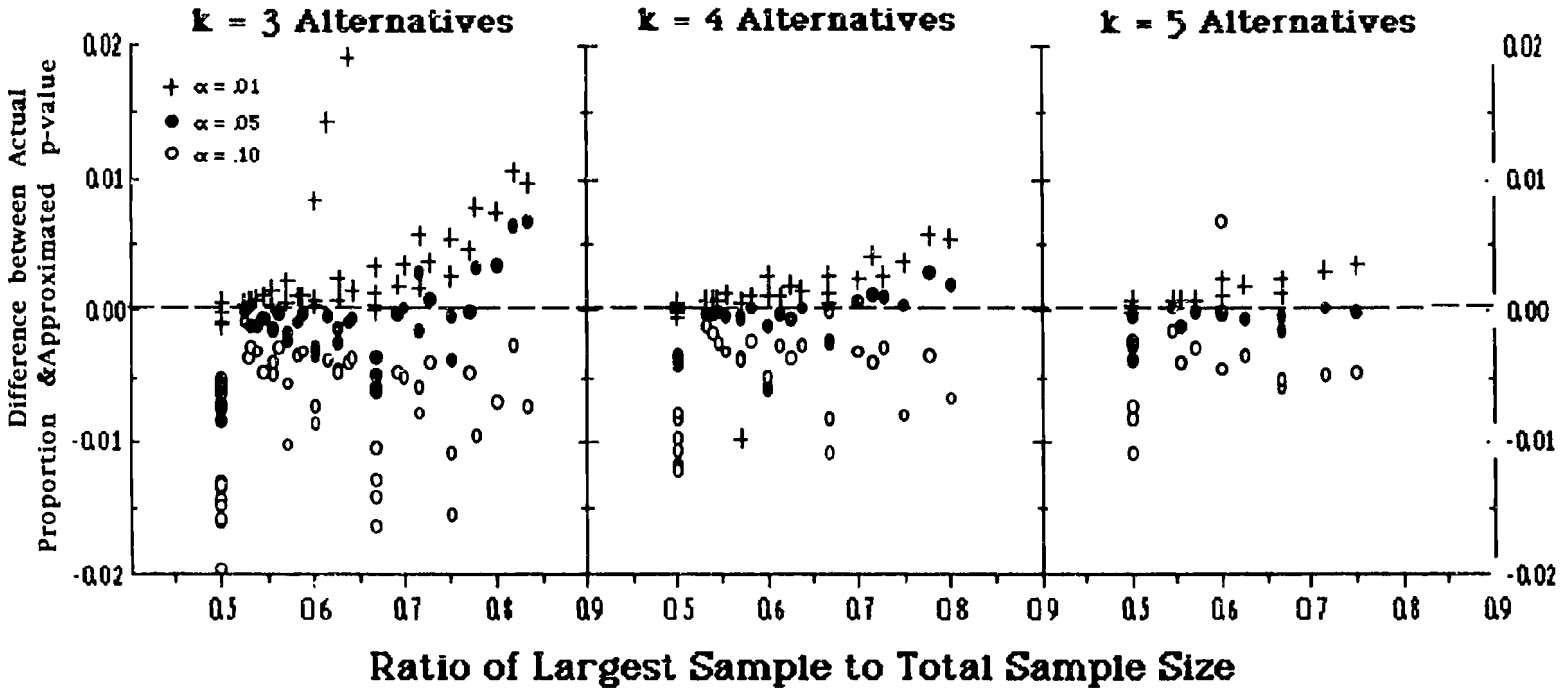

Figure 2. Differences between the actual proportion above the critical value as function of sample size ratio for 3, 4, and 5 alternatives. Horizontal dashed line is at zero which indicates points of perfect fit.

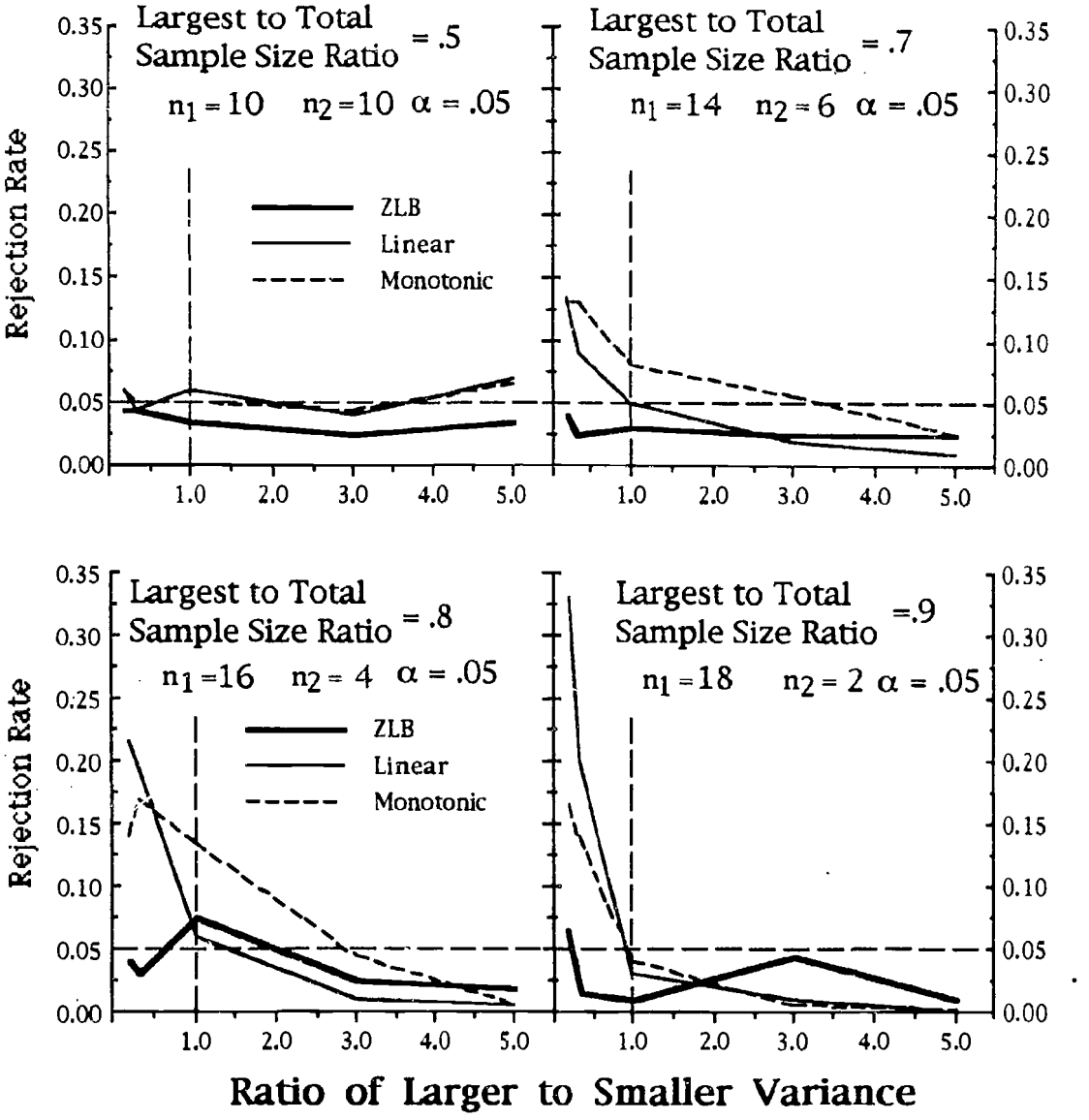

Figure 3. Null hypothesis rejection rates for z_{LB}, linear, and monotonic interaction contrasts as a function of varinace and sample size ratios.

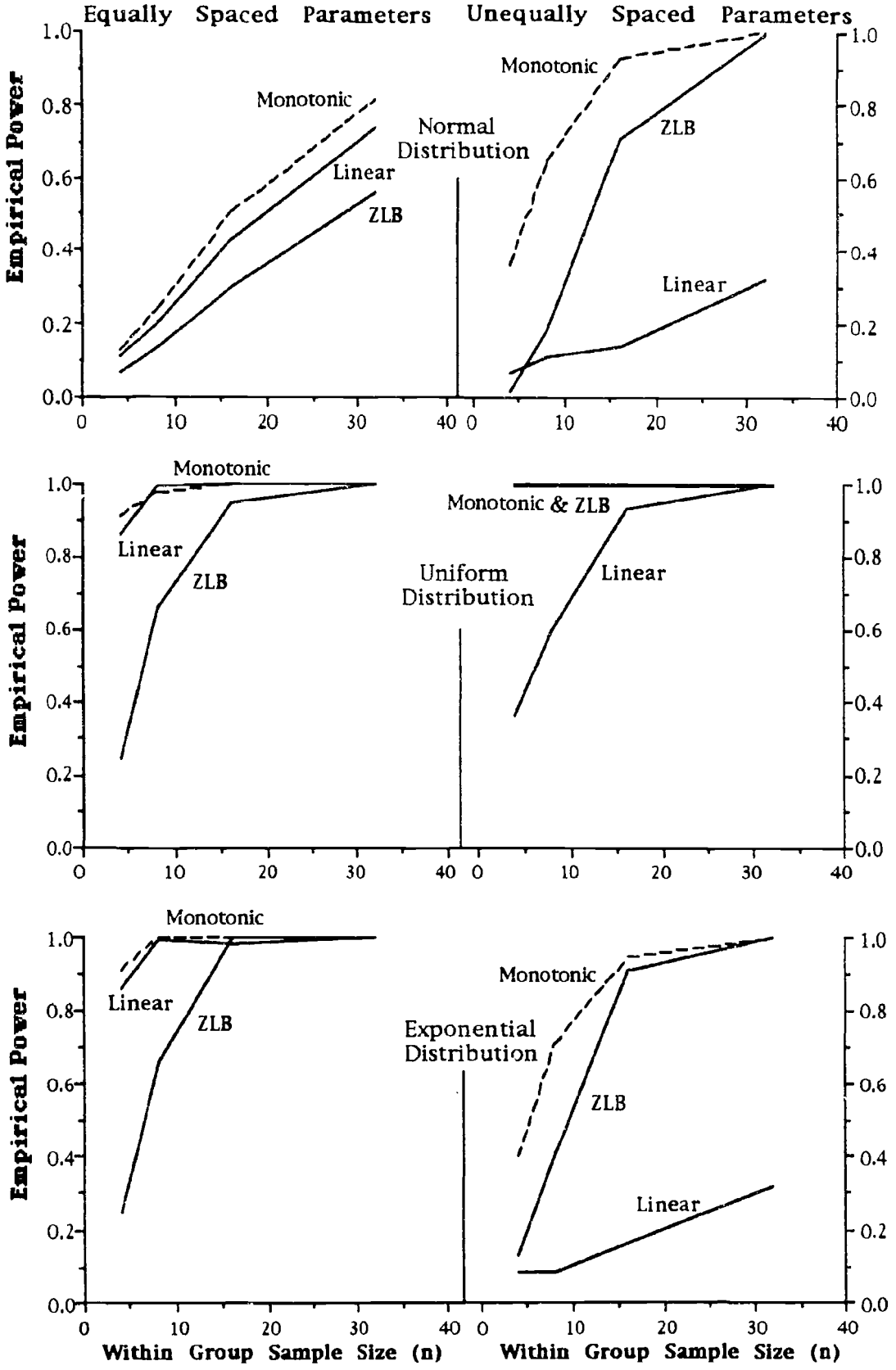

Figure 4. Power as a function of sample size and distribution.